METHOD AND DEVICE FOR DICTIONARY MANAGEMENT, AND DICTIONARY UTILIZATION SYSTEM

Publication number: JP10260984

Publication date: 1998-09-29

Inventor: HIRAKAWA HIDEKI; NOGAMI HIROYASU; SAITO YOSHIMI

Applicant: TOKYO SHIBAURA ELECTRIC CO

Classification:

- international: ***G06F17/28;*** ***G06F17/22;*** ***G06F17/30;*** **G06F17/28; G06F17/22; G06F17/30;** (IPC1-7): G06F17/30; G06F17/22; G06F17/28

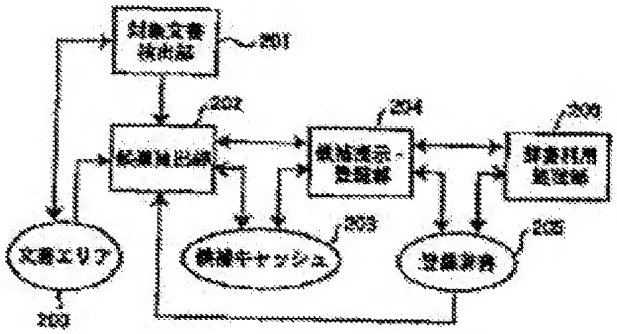
- European:

Application number: JP19970066874 19970319

Priority number(s): JP19970066874 19970319

Abstract of **JP10260984**

PROBLEM TO BE SOLVED: To provide the dictionary management device which can easily support the structuring of a language knowledge dictionary for KANA(Japanese syllabary), machine translation, etc., and the dictionary utilization system that can make natural language processing efficient by using the device. SOLUTION: When a user specifies information identifying existing and nonexisting document data as an object of extraction of a dictionary registration candidate word, an object document detection part 201 detects document data on the basis of the specified information, a candidate extraction part 202 extracts a dictionary registration candidate word from the detected document data, and a candidate display and registration part 204 displays a word selected out of extracted dictionary registration candidate words according to previously specified conditions or the extracted dictionary registration candidate words, thereby registering words that the user selects in a dictionary 205.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－260984

(43)公開日　平成10年(1998)９月29日

(51)Int.Cl.⁶		識別記号	Ｆ Ｉ		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/40	３７０Ｊ
	17/22			15/20	５３２Ｋ
	17/28			15/38	Ｃ
				15/403	３３０Ｃ

審査請求　未請求　請求項の数７　ＯＬ　（全 12 頁）

(21)出願番号　特願平9－66874

(22)出願日　平成９年(1997)３月19日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　平川　秀樹
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　野上　宏康
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　齋藤　佳美
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

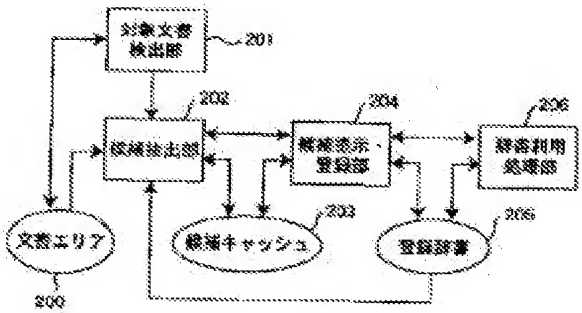
(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】　辞書管理方法および辞書管理装置および辞書利用システム

(57)【要約】

【課題】かな漢字変換や機械翻訳などの言語知識辞書の構築支援が容易に行える辞書管理装置およびそれを用いて自然言語処理の効率化が図れる辞書利用システムを提供する。

【解決手段】ユーザにより、辞書登録候補語の抽出対象となる既存および非既存の文書データを同定する情報が指定されると、対象文書検出部２０１にて、この指定された情報を基に文書データを検出し、候補抽出部２０２にて、この検出された文書データから辞書登録候補語を抽出し、候補提示・登録部２０４にて、この抽出された辞書登録候補語のうち予め指定された条件に基づき選択された語、あるいは、抽出された辞書登録候補語を提示して、これに対しユーザにより選択された語を辞書２０５に登録する。



【特許請求の範囲】
【請求項１】　自然言語処理に用いられる辞書の登録語を管理する辞書管理方法において、
予め指定された前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を基に文書データを検出し、この検出された文書データから前記登録候補語を抽出し、この抽出された登録候補語と予め指定された条件に基づき前記辞書の登録語を更新することを特徴とする辞書管理装置。
【請求項２】　自然言語処理に用いられる辞書の登録語を管理する辞書管理方法において、
予め指定された前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を基に文書データを検出し、この検出された文書データから前記登録候補語を抽出し、この抽出された登録候補語を提示して、この提示された登録候補語に対するユーザによる選択結果に応じて前記辞書の登録語を更新することを特徴とする辞書管理方法。
【請求項３】　自然言語処理に用いられる辞書の登録語を管理する辞書管理装置において、
前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、
この指定手段で指定された情報を基に文書データを検出する検出手段と、
この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、
この抽出手段で抽出された登録候補語と予め指定された条件に基づき前記辞書の登録語を更新する更新手段と、
を具備したことを特徴とする辞書管理装置。
【請求項４】　自然言語処理に用いられる辞書の登録語を管理する辞書管理装置において、
前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、
この指定手段で指定された情報を基に文書データを検出する検出手段と、
この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、
この抽出手段で抽出された登録候補語を提示する提示手段と、
この提示手段で提示された登録候補語に対するユーザによる選択結果に応じて前記辞書の登録語を更新する更新手段と、
を具備したことを特徴とする辞書管理装置。
【請求項５】　辞書の登録語を参照して所定の自然言語処理を行う辞書利用システムにおいて、
前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、
この指定手段で指定された情報を基に文書データを検出する検出手段と、
この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、
この抽出手段で抽出された登録候補語と前記辞書の登録語を参照して、ユーザにより入力された自然言語に対し所定の自然言語処理を行い、前記自然言語の変換候補を生成する自然言語処理手段と、
この自然言語処理手段で生成された変換候補に対するユーザによる選択結果に応じて前記辞書の登録語を更新する更新手段と、
を具備したことを特徴とする辞書利用システム。
【請求項６】　前記更新手段は、ユーザにより選択された変換候補が前記登録候補語であるとき、その登録候補語を前記辞書に登録することを特徴とする請求項５記載の辞書利用システム。
【請求項７】　辞書に登録された語を参照して所定の自然言語処理を行う辞書利用システムにおいて、
所望のサイトとの間を所定の通信回線を介して接続する接続手段と、
この接続手段で接続されたサイトに具備された辞書の登録語を参照して、ユーザにより入力された自然言語に対し所定の自然言語処理を行う自然言語処理手段と、
を具備したことを特徴とする辞書利用システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、例えば、かな漢字変換や機械翻訳など、言語知識辞書を利用して所定の自然言語処理を行う辞書利用システムおよび辞書の構築支援を行う辞書管理装置に関する。
【０００２】
【従来の技術】最近、計算機への日本語の入力を行うＩＭＥ（Ｉｎｐｕｔ　Ｍｅｔｈｏｄ　Ｅｄｉｔｏｒ）や、文書を翻訳する機械翻訳システムなどが広く利用されるようになってきている。この様のソフトウェアは、種々の言語情報を含む辞書を利用している。ユーザは、そのユーザの辞書に、固有の単語を入れることにより、例えば、かな漢字変換の変換精度を向上させることができる。
【０００３】しかし、この辞書への用語の登録は、人手で行うには煩雑であり、これを簡略化するために、文書入力時にユーザからのキー入力や選択処理の情報により、自動的に辞書項目を登録する方式や、既存の文書を指定して、その既存の文書を解析することにより、辞書データを抽出するという方式が考案されてきた。
【０００４】
【発明が解決しようとする課題】ユーザの変換候補の選択など、ユーザからの入力より辞書を作成するという方式では、辞書の学習対象がユーザの入力文書のみであるという点から、学習の範囲が制約されるという問題があ

る。また、既存の文書を指定して、その文書より辞書データを学習するという方式では、この問題を解決できるが、文書をいちいち指定する必要があるため、文書入力を行う際に、必ずしも必要な文書に対してデータの抽出が終了しているというわけでない。このように、従来の技術では、必要なデータを必要となる時までに用意するためのワク組みがない。

【０００５】本発明は、このような辞書登録管理システムの課題を解決するためになされたもので、その目的は、非既存のデータを指定し、データを取得時に辞書データの抽出処理を行っておき、また、登録すべきデータとそうでないデータをそれとなく提示・選択することにより、簡便に辞書データを登録・管理する（辞書の構築支援）環境を提供する辞書管理装置を提供することにある。

【０００６】また、文書ブラウザなど、ネットワーク上のサイトやページを訪れたりする場合には、このサイトより、言語処理に役立つ情報を入手・利用することにより、サイトに応じた処理のカスタマイズ化が可能な自然言語処理装置を提供することを目的とする。

【０００７】

【課題を解決するための手段】本発明の辞書管理方法は、自然言語処理に用いられる辞書の登録語を管理する辞書管理方法において、予め指定された前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を基に文書データを検出し、この検出された文書データから前記登録候補語を抽出し、この抽出された登録候補語と予め指定された条件に基づき前記辞書の登録語を更新することにより、辞書構築支援が容易に行え、自然言語処理の効率化が図れる。

【０００８】また、本発明の辞書管理方法は、自然言語処理に用いられる辞書の登録語を管理する辞書管理方法において、予め指定された前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を基に文書データを検出し、この検出された文書データから前記登録候補語を抽出し、この抽出された登録候補語を提示して、この提示された登録候補語に対するユーザによる選択結果に応じて前記辞書の登録語を更新することにより、辞書構築支援が容易に行え、自然言語処理の効率化が図れる。

【０００９】本発明の辞書管理装置は、自然言語処理に用いられる辞書の登録語を管理する辞書管理装置において、前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、この指定手段で指定された情報を基に文書データを検出する検出手段と、この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、この抽出手段で抽出された登録候補語と予め指定された条件に基づき前記辞書の登録語を更新する更新手段と、を具備したことにより、辞書構築支援が容易に行え、自然言語処理の効率化が図れる。

【００１０】本発明の辞書管理装置は、自然言語処理に用いられる辞書の登録語を管理する辞書管理装置において、前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、この指定手段で指定された情報を基に文書データを検出する検出手段と、この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、この抽出手段で抽出された登録候補語を提示する提示手段と、この提示手段で提示された登録候補語に対するユーザによる選択結果に応じて前記辞書の登録語を更新する更新手段と、を具備したことにより、辞書構築支援が容易に行え、自然言語処理の効率化が図れる。

【００１１】また、本発明の辞書利用システムは、辞書の登録語を参照して所定の自然言語処理を行う辞書利用システムにおいて、前記辞書への登録候補語の抽出対象となる既存および非既存の文書データを同定する情報を指定する指定手段と、この指定手段で指定された情報を基に文書データを検出する検出手段と、この検出手段で検出された文書データから前記登録候補語を抽出する抽出手段と、この抽出手段で抽出された登録候補語と前記辞書の登録語を参照して、ユーザにより入力された自然言語に対し所定の自然言語処理を行い、前記自然言語の変換候補を生成する自然言語処理手段と、この自然言語処理手段で生成された変換候補に対するユーザによる選択結果に応じて前記辞書の登録語を更新する更新手段と、を具備したことにより、自然言語処理の効率化が図れる。

【００１２】また、本発明の辞書利用システムは、辞書に登録された語を参照して所定の自然言語処理を行う辞書利用システムにおいて、所望のサイトとの間を所定の通信回線を介して接続する接続手段と、この接続手段で接続されたサイトに具備された辞書の登録語を参照して、ユーザにより入力された自然言語に対し所定の自然言語処理を行う自然言語処理手段と、を具備したことにより、自然言語処理の効率化が図れる。

【００１３】

【発明の実施の形態】以下、本発明の実施形態について図面を参照して明する。

（第１の実施形態）図１は、本発明の第１の実施形態に係る、自然言語処理に用いられる辞書の登録語を管理する辞書管理装置の構成例を示すブロック図である。

【００１４】図１において、１０１はネットワークよりデータを取り出したり、ネットワークへデータを送信したりする機能を持つネットワーク入出力部である。１０２は各種データを記憶する機能を有するデータ記憶部。１０３はディスプレイなど表示手段を含みユーザへ情報を提示する機能を有するデータ提示部、１０４はキーボード、ペン入力など、ユーザの所望するデータを入力する機能を有するユーザデータ入力部である。１０５はデ

ータ収集の対象となるデータの存在を検出する機能を有する収集対象検出部、１０６は収集対象データより、データ登録候補を生成する候補生成部、１０７は上記各部を所定の動作を行うように制御する制御部である。

【００１５】以下、制御部１０７の動作を説明する。制御部１０７は、ユーザにより指定された抽出対象から登録すべき候補データの収集、提示等を行う辞書管理エージェントに基づき、図１のハードウエア的な各部を制御するようになっている。このエージェントの機能を実現するプログラムは、図１の記憶部１０２に格納されていてもよい。

【００１６】図２は、図１の辞書管理装置の起動時の動作を説明するためのフローチャートである。辞書管理装置の起動時には、まず、収集対象の文書データを特定する情報（特定情報）の指定を行う（ステップＳ１）。

【００１７】図３は、この指定を行うためのインターフェイス画面の例である。図３において、「ウォッチポイント１」は、抽出対象全体に付けられた名前であり、ここで定義された複数のエリアは、この名前で参照・指示される。

【００１８】図３において、「抽出対象」とは、登録すべき候補データを抽出する対象であり、例えば、コンピュータのローカルデータを格納するエリア、ネットワークにおけるデータエリア、特定のアプリケーションの指定するデータエリアを抽出対象ロケーションとして指定することができる。

【００１９】最初に抽出対象として指定された「ｃ：￥￥ｍｙｄｏｃｕｍｅｎｔｓ￥ｒｅｐｏｒｔ￥」は、ローカルのファイルシステムのディレクトリを指定した例である。「下位エリア」は、この「ｒｅｐｏｒｔ」というディレクトリのさらに下位にあるディレクトリを抽出対象に含むか否かを設定するパラメータであり、この場合は、「含まない」を選択しているため、ディレクトリ「ｒｅｐｏｒｔ」に含まれるファイルが抽出の対象となる。また、「対象」の項目において、「＊．ｄｏｃ」が指定されていることにより、このファイル識別子を持つファイルのみが抽出対象となる。ここで、「＊」は、全ての文字列にマッチするワイルドカード文字である。この指定があると、例えば、「ａｂｃ．ｔｘｔ」というファイルは、抽出対象とならない。

【００２０】２番目に抽出対象として指定された「ＮｅｔＳｅｒｖｅｒ１￥￥ｇｒｏｕｐ＿ｄｏｃｕｍｅｎｔｓ￥」は、「ＮｅｔＳｅｒｖｅｒ１」という名前で定義されたネットワーク上に存在し、接続されているコンピュータの有するディレクトリ「ｇｒｏｕｐ＿ｄｏｃｕｍｅｎｔｓ」にある、あるいは、将来的に置かれるファイルを指示している。この場合、「下位エリア」の項目には、「含む」が指定されているため、このディレクトリの下位にあるディレクトリに置かれるファイルも抽出対象となる。さらに、「対象」の項目には、「＊．＊」が指定されており、これは、ファイルの名前とその識別子がいずれもよいという指定になっている。　２番目に抽出対象として指定された「Ｕｄｍａｉｌ－ｆｏｌｄｅｒ」は、アプリケーションソフトの「ｕｄｍａｉｌ」という、メイルシステムのメイルの受け取り口であるディレクトリを指定している。この種のアプリケーションのデータ保存場所に関しては、各種アプリケーションの情報を保存してあるファイルを参照したり、アプリケーションに特定のプロトコルで問い合わせたりすることにより、その情報を得ることができる。

【００２１】図３に指示されるような内容は、システムがあらかじめ用意したデータを、ユーザに提示して選択したり（既存アプリケーションのファイルなど）、例えば、Ｗｉｎｄｏｗｓ（マイクロソフト社の登録商標）のエクスプローラのような表示方法で、ユーザにファイルのディレクトリ構造を示して入力してもらうなどし、これら情報を図１の記憶部１０２へ記憶しておくことにより、随時参照することが可能である。

【００２２】辞書管理装置は、図２のステップＳ１で指定された抽出対象を監視し、辞書データ収集対象が現れると、それに対して辞書登録候補の収集や提示などを行うが、次に、図２のステップＳ２において、その際の振る舞いに関するパラメータなどを設定する。

【００２３】なお、本実施形態に係る辞書管理装置は、ユーザにより指定された抽出対象から登録すべき候補データの収集、提示等を行う辞書管理エージェントを複数定義できるものとする。

【００２４】図４に、パラメータ設定画面の一例を示す。図４において、「ＤｉｃＡｇｅｎｔ＿ＲＳ１」は、辞書管理エージェントの名前である。本実施形態にかかる辞書管理装置は、複数の辞書管理エージェントを定義でき、必要に応じて複数のエージェントを走らせたりすることができる。複数エージェントの起動については、既存ＯＳの複数プロセスの実行の枠組を利用して容易に実現できるので説明は省略する。

【００２５】図４に示した設定画面の「対象」の項目に指定されている「ウォッチポイント１」は、図３で示したデータ抽出の対象となるエリアに対応しており、この「ウォッチポイント１」の部分を指示・クリックすることにより、図３の画面を呼び出し、再設定なども行える。

【００２６】「監視インターバル」の項目には、辞書管理エージェントが「対象」の項目にて指定されたエリアをどのくらいの時間間隔で新規データがないかをチェックにゆくかを設定するようになっている。ここでは、計算機を起動した時にのみ行う「起動時のみ」と、特定の時間インターバルでチェックする２つがあり、後者では、時間を設定する。なお、図４では、監視インターバルとして、「５分」が指定されている。また、監視インターバルは、辞書管理エージェントがチェックを行った

時間を記憶しておき、この時間と現在の時間を比較することにより、容易に実現可能である。

【００２７】「抽出候補」の項目には、対象文書から抽出する登録候補を何にするかを設定するようになっている。「候補キャッシュ」の項目には、抽出された辞書登録候補を記憶する記憶域の記憶容量、例えば、記憶する辞書登録候補の語数を設定するようになっている。これらについては後述する。

【００２８】「自動登録」の項目には、辞書登録候補がファイルなどにより抽出された場合に、これを自動的に辞書に登録するか否かを設定するようになっている。自動登録先は、通常のユーザによる辞書への登録語と同様に扱ってもよいが、ここでは、自動登録用の辞書を別途用意する、あるいは、自動登録されたという識別情報を入れて登録する。これにより、ユーザが辞書を編集するなどの際に容易に自ら登録した語か否かを判定することができる。

【００２９】なお、一般的に、登録した辞書を利用して言語処理（例えば、かな漢字変換）するソフトウェアにおいて、辞書の登録語をユーザにより選択／利用（例えば、かな漢字変換で登録された語を同音語選択するなど）された際にその情報を記憶しておくことができる。この情報により、ユーザが利用したデータのみを残して、他の使用されなかった単語を削除することができる。さらに、登録した単語にタイムスタンプ（登録された時間）を押せば、所定の時間がたった後に使用されなかった単語を手動もしくは自動で削除することができ、これにより登録語が多くなり処理効率が劣化するのを防止することが可能である。

【００３０】本実施形態では、登録語のタイムスタンプでなく、候補キャッシュを利用しているので、ユーザにより指定された所定の候補数までしか記憶せず、それ以上になる場合には、（利用されなかった）古い候補を削除してしまうという方法を採用しておりメモリ効率的には優れている。

【００３１】図４の説明に戻り、「自動登録」の項目において、「有り」を選択した場合には次に登録の条件を設定する。「登録条件」の項目には、自動登録される単語が満たすべき条件を設定するようになっていて、登録する単語の精度（有用性）をあげるために利用される。

【００３２】登録条件としては、例えば、「最低頻度」、「確信度」がある。「最低頻度」は、対象文書から抽出された候補の頻度に関する条件であり、図４では、頻度が２以上のものを自動登録とする指定となっている。頻度の詳細については、後述する。「確信度」は、登録候補抽出において、その候補の尤もらしさに対応する点数である。その詳細は後述するが、図４では、確信度「１」を設定している。

【００３３】「登録候補自動表示」の項目には、抽出された辞書登録候補を図１の提示部１０３に提示するか否かを指定するようになっている。図４に示した表示画面の下部にある「実行」、「キャンセル」のボタンは、この設定処理の実行およびキャンセルを行うためのものである。

【００３４】次に、辞書登録候補語を抽出する対象文書からの辞書情報の獲得処理に関する辞書管理資源の動作、すなわち、各辞書管理エージェントの処理動作について説明する。

【００３５】図５は、辞書管理エージェントの処理の流れを説明するための概念図である。図５において、対象文書検出部２０１は、例えば、図４の設定画面から設定された監視インターバルに基づいて、図３で示したような画面から設定された抽出対象文書データの特定情報を基に、抽出対象の文書エリア２００に新規抽出対象の文書データが存在するか否かを確認する。

【００３６】新規文書データであるか、既に候補抽出を行った文書データであるかは、そのファイルの作成された時間を比較することにより行われる。これは、文書データファイルの収集を開始した時間ｔ１を所定のメモリに記憶し、前回に抽出を開始した時間ｔ２（メモリに記憶されている）と、ファイルの作成時間ｔ３とが、ｔ１＜ｔ３≦ｔ２の場合に、作成時間の新しい文書データファイル、すなわち、抽出対象ファイルであるという判定をする。指定されたディレクトリなどのファイルが走査され、作成時間の新しい文書データファイルが抽出される。この際に、図３の「下位エリア」の項目で指示された下位エリアを走査する処理、あるいは「対象」の項目で指示された対象を限定する処理が行われ、所定の条件に合うファイルの名前が収集される。最後に収集の開始の時間ｔ２が記憶され、これは次回の抽出時に利用される。このようにして収集された結果、得られる抽出対象文書データのファイルリストは、候補抽出部２０２に渡される。

【００３７】候補抽出部２０２は、渡されたファイルリストのおのおののファイルより、辞書登録候補を抽出し、これを候補キャッシュ２０３へ出力する。図６は、候補抽出部２０２により抽出され、候補キャッシュ２０３に格納された辞書登録候補レコードの一例を示したものである。図６において、セミコロンで区切られたおのおのがフィールドは、読み、綴り、品詞、確信度、頻度である。辞書登録候補レコードは、候補キャッシュ２０３に基本的に候補抽出部２０２の出力する辞書登録候補に対して、ＦＩＦＯ（Ｆｉｒｓｔ　ｉｎ　Ｆｉｒｓｔ　ｏｕｔ）で記憶されている。候補キャッシュ２０３の大きさは、図４の「候補キャッシュ」の項目で設定される語数（この場合３０００）で決定する。

【００３８】図６では、番号「１」の「インターネット」が最も新しい辞書登録候補であり、番号「３０００」が最も古い辞書登録候補である。この状態で、候補キャッシュ２０３に存在しない候補が登録されると、番

号「３０００」の候補「きゃば」は捨てられることになる。

【００３９】候補抽出部２０２から出力される辞書登録候補が候補キャッシュ２０３に存在する候補の場合は、その候補の頻度が「１」加算され、その候補レコードは、キャッシュの先頭に置かれる。この際、優先度は高い方が選択される。

【００４０】候補レコードの「確信度」は、候補抽出処理において付与される。この値は、候補抽出に使用されるパターンもしくは選択に基づくルールにより決定することができる。例えば、漢字２文字あるいはカタカナ列が句点と格助詞に挟まれているときは、最も確信度の高い「Ａ」をアサインする。具体的には、「…、インターネットに…」という文字列から「インターネット」が辞書登録候補として抽出された場合には、その確信度は「Ａ」であり、図６の番号「２」に対応する候補レコードが生成される。また、「する」の前の２文字の漢字列は、確信度「Ｃ」とするという場合には、例えば、「…外部定検した。」という文字列から、「定検」が取り出され、図６の番号「５０」の候補レコードのようになる。なお、「定検」の場合、「定」には、「てい」「じょう」の２つの読みが一般的であるため、「検」の読みである「けん」と組み合わせて「ていけん、じょうけん」の２つが読みの候補として生成されている。

【００４１】次に、図５の候補提示・登録部２０４の処理動作について説明する。例えば、図４で「自動登録」の項目が「有り」に設定されていると、候補提示・登録部２０４は、候補キャッシュ２０３の先頭に追加された候補に対して、チェックを行い、その候補が「自動登録」の「登録条件」を満たしていた場合は、それを登録辞書２０５へ登録する。登録条件のチェックは、辞書登録候補レコードの頻度フィールドや確信度フィールドを参照することにより可能である。

【００４２】図４に示した設定画面において、「登録条件」の項目に変更があった場合には、候補キャッシュ２０３に格納されている変更前の登録条件に該当する候補を提示し、ユーザに対して登録を行うか否かの問い合わせを行った後、必要であれば登録辞書２０５に対する登録処理を行う。

【００４３】登録辞書２０５に登録された語は、候補抽出部２０２により利用されるため、登録された以降は、その単語自体は、未知語とならないため、登録候補として抽出されることは通常ない。候補提示・登録部２０４が候補を登録する登録辞書２０５は、あらかじめユーザが指定するか、あるいは、デフォルトとなる辞書を利用するが、自動登録された語に対しては、その旨の情報を付与して他と区別できるようにする。

【００４４】候補提示・登録部２０４は、上記登録処理の他に、候補提示処理を行う。これは、例えば、図４における「登録候補自動表示」の項目を「有り」に設定してある場合に実行される。候補提示・登録部２０４は、候補キャッシュ２０３に辞書登録候補があると、これをユーザに提示する。

【００４５】図７は、候補キャッシュ２０３に格納された辞書登録候補の提示画面の一例を示したものである。図７（ａ）において、「ＤｉｃＡｇｅｎｔ＿ＲＳ１」は、辞書管理エージェントの名前であり、これがどのエージェントにより抽出されているかを示す。右上部に表示される「５／２５」は、全体の辞書登録候補が「２５」あり、このうち５番目を表示していることを示している。「インターネット」以下は、各辞書登録候補の表示例である。

【００４６】この辞書登録候補は、時間とともに表示が変化し、例えば、図７（ａ）の画面は、図７（ｂ）の画面のように変化する。図７（ｂ）では、６番目からの単語が表示されている。単語の順番は、候補キャッシュ２０３にＦＩＦＯを用いているので、新しい候補が１番目、古い候補ほど大きい番号となる。単語の提示の順番については、この他に頻度の順に提示するなども考えられる。

【００４７】他の辞書登録候補の表示方法として、例えば、電光掲示板形式で流すなどの方法も考えられる。図８は、その様子を示したもので、辞書登録候補の各単語は、右から左へ流れてゆく。この方法の方が表示領域が狭く、他の仕事をしている最中にでも流しておき、時間のあいた時に登録処理などを行うというように使用することができる。

【００４８】図７、図８に示すように、提示される辞書登録候補の文字列の後には、括弧で付加情報がつく。基本的には、登録候補の後の括弧の中には、読みのリストと「×」が表示される。カタカナ書き語の辞書登録候補の場合は、読みの代わりに「○」を表示する。この「○」をマウスでクリックすると、その候補が登録候補として認定されたと解釈する。また、複数の読みのうち特定の読みをクリックすると、その読みでの登録が指示されたと判定する。例えば、「定検」に対して「ていけん」をクリックすると、「ていけん」という読みで「定検」を登録するというように解釈される。

【００４９】「×」をクリックした場合は、その候補は登録しないというように判断される。登録しないと判断された候補は、特定の領域（図示しない）にネガティブな例として記憶し、候補抽出部２０２で候補生成の際にそれの生成を行わない、もしくは候補提示・登録部２０４で登録候補として表示しないなどの仕組みをいれることができる。この場合、ユーザの「×」の指示自体が誤っていたりすると、１回の削除操作で以降、候補として選択することができなくなるので、複数回削除すると候補としての表示をやめたり、以降候補として生成しない旨のメッセージを出し、ユーザの確認を得るなどの機能を付与することにより、こうした不具合は、最小限にお

きることができる。また、Ｗｉｎｄｏｗｓ９５（マイクロソフト社登録商標）の「ゴミ箱」のように、「×」により削除されたデータは、画面表示からは消すが、別の領域（図示しない）に保存しておき、復元が可能なようにしておくという構成にすることも容易に可能である。

【００５０】ユーザが、図７あるいは図８のように提示された辞書登録候補の付加情報のうち、「○」あるいは読みをクリックした際には、例えば、図９に示すような辞書登録画面が表示される。例えば、「定検」に対して「ていけん」あるいは「○」がクリックされると、図９に示すように、見出し「定検」、読み「ていけん」の他に推定される品詞が表示され、ユーザは、この画面を見て必要に応じて、この登録データを修正することもできる。

【００５１】図９において、登録ボタンは辞書の登録を、削除ボタンは候補表示からの削除を、キャンセルは、登録操作のキャンセルをそれぞれ行うためのものである。なお、図５において、文書エリア２００、候補キャッシュ２０３、登録辞書２０５は、図１の記憶部１０２の記憶領域内に設けられることができる。また、辞書利用処理部２０６は、例えば、かな変換処理、機械翻訳等の登録辞書２０５を用いて所定の自然言語処理を実行する言語処理エージェントである。

【００５２】以上、説明したように、上記第１の実施形態によれば、ユーザにより、辞書登録候補語の抽出対象となる既存および非既存の文書データを同定する情報（図３参照）が指定されると、対象文書検出部２０１にて、この指定された情報を基に文書データを検出し、候補抽出部２０２にて、この検出された文書データから辞書登録候補語を抽出し、候補提示・登録部２０４にて、この抽出された辞書登録候補語のうち予め指定された条件に基づき選択された語、あるいは、抽出された辞書登録候補語を提示して、これに対しユーザにより選択された語を辞書２０５に登録等することにより、ユーザ固有の文書から、言語処理のための辞書候補情報を、自動的に抽出しておき、その辞書データを利用することが可能となり、言語処理システムの効率化が行える。また、得られた辞書登録候補情報は、ユーザが自然に選択することにより、テンポラルなものからそうでないものへと変換することが可能であり、適切なデータの構築が容易に行える。

【００５３】また、辞書抽出対象の文書データが新たに検出されたり、ある文書データからの辞書登録候補の抽出が進むにつれて、それらを順次提示したり、この提示される辞書登録候補に対して、登録／削除／変更の指示を与えて辞書登録候補の提示態様を変化させることにより、辞書構築の支援が容易に行える。

【００５４】このように、ユーザにより辞書登録候補の抽出元の文書データを特定する情報が与えられると、その特定情報を基に文書データを検出して、その後、この対象より自動的に辞書データを抽出しておくことが可能となり、適切な時点でユーザの辞書登録・管理が行え、かな漢字変換などでユーザが辞書を使用する際には、既に辞書登録が済んでおり、文書作成効率が向上するという効果がある。

【００５５】（第２の実施形態）次に、第１の実施形態で説明した、既存、非既存の文書データから抽出された語を登録して登録辞書生成する辞書管理装置を適用した辞書利用システムについて説明する。ここでは、一例として、かな漢字変換辞書に関する説明を行うが、本発明は、機械翻訳、検索など、辞書を利用するシステム全般に対して適用可能であることは言うまでもない。

【００５６】図１０は、第２の実施形態に係る辞書利用システムの構成例を示したものである。図１０において、言語処理エージェント３０３は、既存技術などによるかな漢字変換を行う機能を有し、入力部よりかな文字列を入力し、基本的に登録辞書中のかな漢字変換データを利用して、かな漢字混じり列に入力文字列を変換して、提示部に提示する制御を司るものである。

【００５７】候補生成エージェント３０１は、第１の実施形態で説明したような、文書などから辞書情報を抽出し、辞書登録候補を登録候補キャッシュ３０２に登録する制御を司るものである。

【００５８】第２の実施形態に係る辞書利用システムでは、登録候補キャッシュ３０２にテンポラルに登録された辞書情報も言語処理エージェント３０３で処理に利用し、この結果、言語処理エージェント３０３より得られた出力を提示部３０４に提示し、この情報の中からユーザが適切なものを選択し、その情報が入力部３０５より得られた場合に、言語処理エージェント３０３は、選択された情報を登録候補キャッシュ３０２から登録辞書３０６に移動する、あるいは、それに準じる処理をして、他の登録候補とは異なった形態に該登録候補を変形する処理を行うようになっている。

【００５９】なお、候補生成エージェント３０１、言語処理エージェント３０３の機能を実現するプログラムは、所定のメモリに格納され、ＣＰＵがこのプログラムに基づき提示部３０４、入力部３０５、および登録候補キャッシュ３０２、登録辞書３０６を格納するメモリにアクセスするなどの所定の処理動作を実行するようになっていてもよい。

【００６０】以下、具体例を用いて説明する。登録候補キャッシュ３０２には、例えば、図６に示した形式と内容の辞書登録候補レコードが格納されるとする。

【００６１】登録辞書３０６に、例えば、「ヒっこう：実行：サ変」という辞書項目が登録されているとする。この状態において、ユーザより、入力部３０５を介して「いんたーねっとふうさでていけんをヒっこうする」という入力が言語処理エージェント３０３に送られてき

たとする。

【００６２】言語処理エージェント３０３は、この入力された文字列に対して、少なくとも、登録辞書３０６と登録候補キャッシュ３０２の２つを用いて辞書引きを行い、この結果に対して形態素解析などを行い漢字候補を生成する。このような、かな漢字変換処理については、既存技術を用いることで実現可能であり、詳細な説明は省略する。

【００６３】この場合、「インターネットブラウザ」が登録候補キャッシュ３０２に予め記憶されているので、かな漢字変換処理の第一次出力結果としては、「インターネットブラウザで定型を実行する」が得られ、提示部３０４に提示される。この段階では、提示された候補は、未選択の状態であり、ユーザによる選択が可能である。

【００６４】出力結果である文字列の「インターネットブラウザ」の部分には、ひらがな書き語「いんたーねっとぶらうざ」などが候補として存在している。ユーザは、提示されている候補が正しい候補であるので、「インターネットブラウザ」に対して選択・確定キーを押すなどして、選択を行う。

【００６５】この操作を検知すると、言語処理エージェント３０３は、選択された候補が登録候補キャッシュ３０２の辞書項目より生成されたものであるということを判定し、そうである場合は、この辞書項目を登録候補キャッシュ３０２から登録辞書３０６へ移し、いわゆる辞書登録を行う。あるいは、登録候補キャッシュ３０２中の候補にユーザ選択された旨の情報を付与しておく。これにより、後の時点でユーザに登録の可否を確認しながら辞書登録することもできる。

【００６６】ユーザにより入力された文字列中の「ていけん」の部分では、登録辞書３０６に存在した「定見」と登録候補キャッシュ３０２に存在した「定検」などが選択の候補として保持されている。このとき、ユーザが「定検」を選択した場合には、上記と同じような処理がなされ、「定見」の辞書項目が登録辞書３０６に移動される。この際に、登録候補キャッシュ３０２中では、「定検」に対して、「ていけん」「じょうけん」の２つが読みの候補となっているが、ユーザの入力した文字列が「ていけん」であるため、「じょうけん」の部分は削除される。これにより、「ていけん」という読みに対して「定検」が選択された後は、「じょうけん」という読みに対して、「定検」という変換候補を出力することはなくなる。

【００６７】このように、ユーザ選択に応じて、登録辞書３０６の情報を変化させることにより、適切な候補以外の候補の出力を制限することが可能である。また、こうした、複数の読み候補を持つ語に対しては、非常に的はずれな変換結果が出る可能性があるため、他の一般辞書などに同音語があったら、ユーザ選択がなされるまでは第一候補としては表示しないなどの工夫により、比較的安全に辞書知識の取り組みを行うことが可能である。

【００６８】機械翻訳においても、かな漢字変換と同様な処理が可能となる。機械翻訳においては、対訳文書からの辞書の作成技術が開発されており、対応する対訳文書から、辞書候補を作成することが可能である。例えば、ＣＯＬＩＮＧ９４のプロシーディングにある「Ｂｕｉｌｄｉｎｇ ａｎ ＭＴ Ｄｉｃｔｉｏｎａｒｙ Ｆｒｏｍ Ｐａｒａｌｌｅｌ Ｔｅｘｔｓｂａｓｅｄ ｏｎ Ｌｉｎｇｕｉｓｔｉｃ ａｎｄ Ｓｔａｔｉｓｔｉｃａｌ Ｉｎｆｏｒｍａｔｉｏｎ，ｐｐ７６」では、対訳テキストから翻訳用辞書の作成について論じている。こうした技術により抽出可能な辞書候補も例えば訳語に関して曖昧性を有し、ユーザの訳語の選択などにより、かな漢字辞書と同様な枠組で処理することが可能である。

【００６９】以上説明したように、上記第２の実施形態によれば、ユーザにより指定された辞書登録候補語の抽出対象となる既存および非既存の文書データを規定する情報（図３参照）に基づき、候補生成エージェント３０１にて、文書データを検出して、辞書登録候補語を抽出し、登録候補キャッシュ３０２に格納しておき、言語処理エージェント３０３では、ユーザにより入力された自然言語に対し登録候補キャッシュ３０２に格納された辞書登録候補語と辞書３０３に登録された語を参照して所定の自然言語処理を行って自然言語の変換候補を生成し、この生成された変換候補に対するユーザによる選択結果に応じて辞書３０３の内容を更新することにより、ユーザ固有の文書から、言語処理のための辞書候補情報を、自動的に抽出しておき、その辞書データを利用することが可能となり、言語処理システムの効率化が行える。また、得られた辞書登録候補情報は、ユーザが自然に選択することにより、テンポラルなものからそうでないものへと変換することが可能であり、適切なデータの構築が容易に行える。

【００７０】（第３の実施形態）次に、辞書に登録された語を参照して所定の自然言語処理を行う辞書利用システムが、所定のネットワークを介して接続されたサイトと互いに通信を行って、このサイトに具備される辞書を参照して所定の自然言語処理を行う場合ついて説明する。すなわち、言語解析などに利用する辞書を具備したサイトエージェントと所定のネットワークを介して互いに通信を行って、言語処理に役立つ情報を入手・利用することにより、サイトに応じた言語処理のカスタマイズ化が可能な言語処理エージェントについて説明する。

【００７１】図１１は、第３の実施形態に係る辞書利用システムの構成例を示したものである。ここでの言語処理エージェントのタスクとしては、例えば、かな漢字変換を想定している。すなわち、入力部４０４からのかな・英数字入力は、辞書４０７を利用して、言語処理エー

ジェント４０７により、かな漢字混じり文字列に変換され、提示部４０３に提示される。

【００７２】言語処理エージェント４０２とサイトエージェント４０１は、所定のネットワークを介して互いに情報のやり取りができるようになっている。このチャネルの設定は、図示されていないが、ユーザの指示により、ネットワーク上のサイトを訪問することにより行われる。これは、例えば、ＷＷＷブラウザでインターネット上のＷｅｂページを表示することなどに相当し、ユーザが訪問した時にページに関連して置かれたサイトエージェント４０１と言語処理エージェント４０２が特定のプロトコルでやり取りをして情報伝達をすることが現状の技術でできる。ユーザがＷｅｂページのハイパーリンクをたどって別のＷｅｂページに行くことにより、言語処理エージェント４０２は、別のサイトエージェント４０１と情報のやり取りを行うようになる。

【００７３】なお、サイトエージェント４０１、言語処理エージェント４０２の機能を実現するプログラムは、それぞれ、例えばパーソナルコンピュータ等の端末装置内の所定のメモリに格納され、ＣＰＵがこのプログラムに基づきサイト情報４０５、サイト辞書４０６を格納するメモリ、あるいは、辞書４０７を格納するメモリ、提示部４０３、入力部４０４などにアクセスして所定の処理動作を実行するようになっている。

【００７４】図１１において、サイト情報４０５は、ＷＷＷページの表示情報などに相当する情報であり、サイト辞書４０６は、言語解析などに利用する辞書である。図１２は、ユーザ側端末装置において表示されるサイトの画面表示例を示したものである。このページは、「リーダースワインセラー」というワイン販売の会社を想定しており、「会社概要」、「ワインリスト」、「ワインあれこれ」などは、それぞれのページへのハイパーリンクになっている。このサイトには、「お話窓」というところを通して、サイトのエージェントとコミュニケーションができる。ユーザは、「お話窓」に文を入力すると、それが解析され、それに応じた応答が生成され、ユーザに提示されるようになっている。

【００７５】図１３は、言語処理エージェント４０２のかな漢字変換処理の手順の概要を示したフローチャートである。例えば、図１２に示したようなページがユーザ側の端末装置に表示されている状態（すなわち、ユーザ側端末と所望のサイトとの間に通信回線が設定された状態）で、ユーザは、「お話窓」の領域内にキーボード等の入力部４０４を介して文字列の入力を行う（ステップＳ１１）。このとき、ローマ字入力の場合は、ひらがな文字列に変換する。今、「とらじゃわいんはありますか」という文字列をユーザが入力した場合を例にとり説明する。

【００７６】次に、言語処理エージェント４０２は、辞書データの要求コマンドとともに、このひらがな列をサイトエージェント４０１に送出する。これを受けるとサイトエージェント４０１は、このひらがな列に含まれる可能性のある単語の集合をサイト辞書４０６より検出し、それを辞書データとして要求元の言語処理エージェント４０２に送出する（ステップＳ１２）。言語処理エージェント４０２が受け取る辞書データは、例えば、次のようなもので、１語毎に読み、綴り、品詞がそれぞれ含まれている。

【００７７】とらじゃわいん；トラジャワイン；固有名詞；／とらじゃ；トラジャ；固有名詞；／

言語処理エージェント４０２は、サイトエージェント４０１から送られてきた辞書データと、独自の持つ辞書４０７の情報を利用して、かな漢字変換処理を行う（ステップＳ１３）。そして、この変換結果をユーザに提示して、ユーザにより正解候補の選択が行われる（ステップＳ１４）。

【００７８】この場合、「トラジャワイン」がサイトエージェント４０１より提供されるため、かな漢字変換結果は、「トラジャワインはありますか」となり、正確な変換が可能となることにより、コミュニケーションをより円滑に行うことができる。

【００７９】また、この例では、辞書データを伝達するという構成にしているが、かな漢字変換処理の一部をサイトエージェント４０１に代行させてしまい、その結果を受け取るような構成も可能であるし、また、かな漢字変換処理の大半をサイトエージェント４０１側で行い、言語処理エージェント４０２側から特定の情報を送るというような構成にすることも可能である。

【００８０】また、辞書データではなく、そのサイトに含まれる文書の情報を利用することにより、そのサイトの文書に現れる文字列候補を優先するという構成も可能である。例えば、「トラジャワイン」が、そのページに含まれている場合には、「とらじゃ」という入力に対して、「虎じゃ」ではなく、「トラジャ」を優先して出力する。これは、基本的に文書を文字列スキャンすることにより可能である。こうした手法を用いた場合には、その文書だけでなく、その文書の関連文書、例えば、ハイパーリンクでつながっている文書を対象とするなどが考えられる。この方法の利点は、サイト側に特殊なエージェントを想定する必要がないことである。特にＷＷＷの場合は、空き時間でリンク先のページをプリフェッチすることにより、リンクをたどる操作が早くなるという利点がある。文字データだけのプリフェッチは、比較的高速に実行可能である。

【００８１】また、かな漢字変換処理を高速に行うためには、１文字が入力されるごとに情報を送って、入力と同時平行的にサイトエージェント４０１側で辞書検索を行うなどの工夫が考えられる。

【００８２】以上により、個々のサイトに応じた入力の適応が実現できる。以上説明したように、上記第３の実

態形態によれば、所望のサイト（サイトエージェント４１）との間を所定の通信回線を介して接続し、言語処理エージェント４０２では、ユーザにより入力された自然言語に対し、所定の通信回線を介して接続されたサイトに具備された辞書４０６に登録された語を参照して所定の自然言語処理を行い、自然言語の変換候補を生成することにより、各サイトに固有の辞書データを利用することができるため、サイトなどの環境に応じた言語処理のカスタマイズが可能となり、自然言語処理の効率化が図れる。

【００８３】なお、本発明は上記第１～第３の実施形態に限定されるものではなく、翻訳、検索、音声入力など種々の言語処理アプリケーションに適用することができる。また、英語やフランス語など、任意の言語に対して適用することが可能である。要するに、本発明の主旨を逸脱しない範囲で種々変形して実施することができる。

【００８４】

【発明の効果】以上説明したように本発明によれば、ユーザ固有の文書から、言語処理のための辞書候補情報を、自動的に抽出しておき、その辞書データを利用することが可能となり、言語処理の効率化が行える。

【００８５】また、得られた辞書候補情報は、ユーザが自然に選択することにより、テンポラルなものからそうでないものへと変換することが可能であり、適切なデータの構築が容易に行える。

【００８６】さらに、計算機ネットワークに存在する各種サイトに接続した場合には、そこに固有のデータを利用することができるため、サイトなどの環境に応じた言語処理のカスタマイズが可能となり、言語処理の効率化が図られる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に係る辞書管理装置の構成例を示した図。

【図２】図１に示した辞書管理装置の起動時の動作を説明するためのフローチャート。

【図３】収集対象文書データを特定するための情報を指定するインターフェイス画面の一例を示した図。

【図４】辞書管理エージェントの振る舞いに関するパラメータを設定する設定画面の一例を示した図。

【図５】辞書管理エージェントの処理の流れを説明するための概念図。

【図６】候補キャッシュに格納された辞書登録候補レコードの一例を示した図。

【図７】辞書登録候補の提示画面の一例を示した図。

【図８】電光掲示板形式で辞書候補情報を提示する場合の提示例を示した図。

【図９】辞書登録画面の一例を示した図。

【図１０】本発明の第２の実施形態に係る辞書利用システムの構成例を示した図。

【図１１】本発明の第２の実施形態に係る辞書利用システムの構成例を示した図。

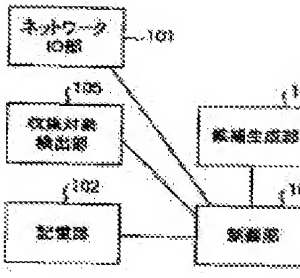
【図１２】サイトの画面表示例を示した図。

【図１３】図１１の言語処理エージェントのかな漢字変換処理の概要を示したフローチャート。

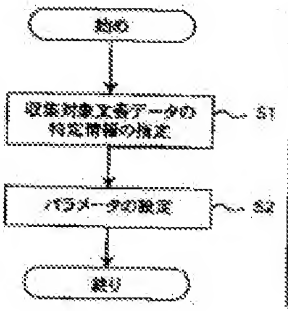
【符号の説明】

１０１…ネットワーク入出力部
１０２…データ記憶部
１０３…データ提示部
１０４…データ入力部
１０５…収集対象検出部
１０６…候補生成部
１０７…制御部。

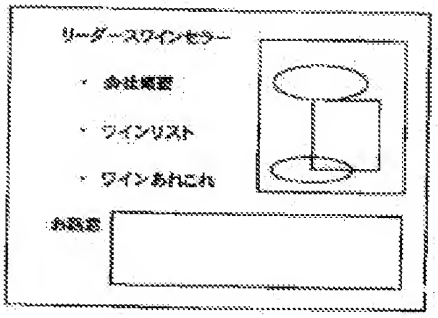
【図１】



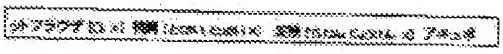
【図２】



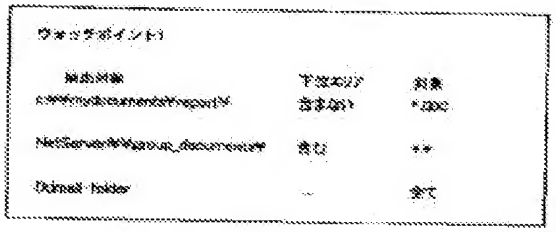
【図１２】



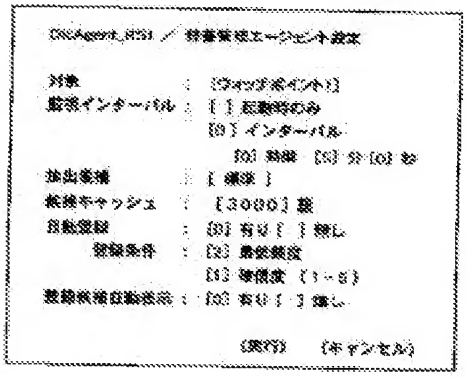
【図８】



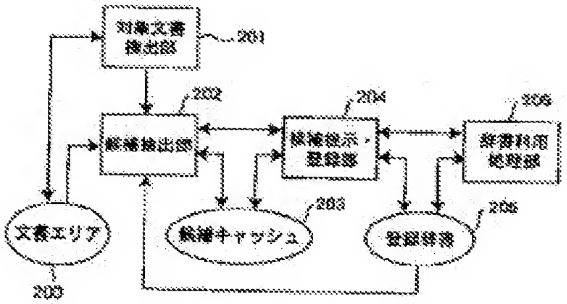
【図3】



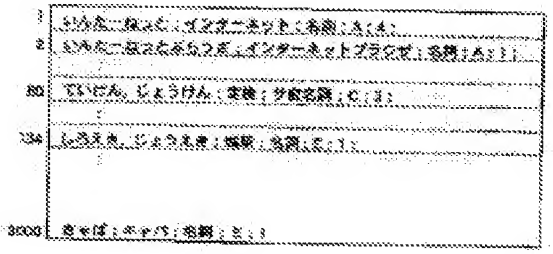
【図4】



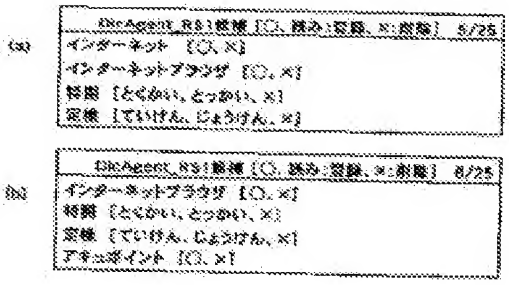
【図5】



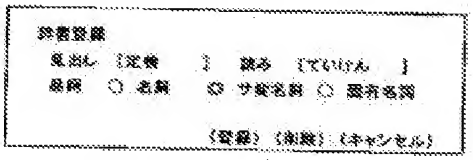
【図6】



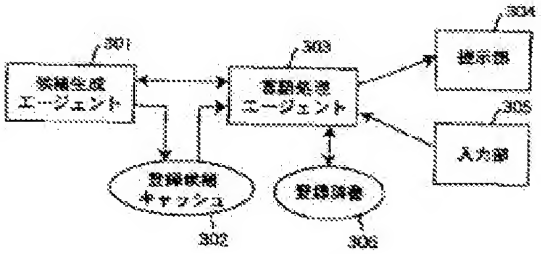
【図7】



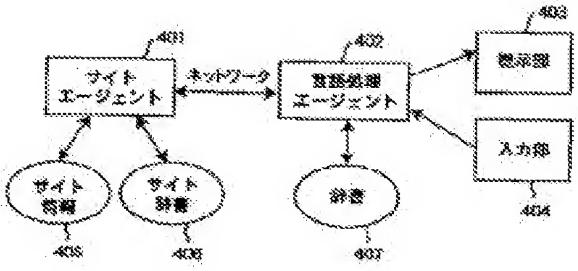
【図9】



【図10】



【図11】



【図13】